ОТБОР ПОЛЕЗНЫХ ПРИЗНАКОВ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ НА БАЗЕ КАРТЫ КОХОНЕНА

A. Pcholkin

Computer Science Department *University of Latvia*
I9 Raina Blvd., Riga LV-1586, LATVIA
Phone: +371 7228611

Рассматривается модификация правила обучения Хеббиана для отбора полезных признаков в задаче кластеризации многомерных данных. Модифицированный алгоритм не только группирует объекты по схожести описаний, но и автоматически производит отбор полезных признаков для улучшения качества группирования. Проводится серия сравнительных экспериментов, в которых модифицированный алгоритм оказывается более работоспособным, чем сеть Кохонена, которая рассчитана на предварительный отбор полезных признаков.

Ключевые слова: карта Кохонена, кластеризация, правило обучения Хеббиана, нейронные сети, самоорганизация.

Введение

Самоорганизующаяся нейронная сеть Кохонена [3,4,5] обучается на примерах без учителя и рассчитана на предварительный отбор признаков до начала обучения. Последнее обстоятельство создает множество трудностей при практическом использовании данной сети. Карта Кохонена не способна улучшить качество кластеризации [1,2] за счет отбора полезных признаков, т.к. не производит ранжировку входных сигналов по степени их важности. По этой причине представляет интерес исследование влияния предварительного отбора полезных признаков на повышение качества кластеризации.

Постановка задачи

Алгоритмы [1], группирующие многомерные данные, исходя из схожести описаний, рассчитаны на предварительный отбор полезных признаков (под понятием “полезный признак” в данном случае понимается такой признак, при использовании которого улучшится качество кластеризации). Данные алгоритмы, как правило, ведут поиск устойчивых сочетаний значений признаков, но не ранжируют их по степени важности.

В данных алгоритмах предполагается, что признаки подобраны достаточно удачно, и объекты хорошо группируются во всем пространстве признаков. Однако, это условие становится практически неприменимым при работе с данными большой размерности. В этом случае логичнее предположить, что в пространстве признаков имеются подпространства (подмножества признаков), в которых объекты хорошо группируются.

Так как сеть Кохонена основана на правиле обучения Хеббиана [4], то учитывая вышесказанное, представляет интерес решение задачи кластеризации объектов, имеющих большое количество признаков без их предварительного отбора. Указанную задачу предлагается решить путем модификации правила обучения Хеббиана. Суть модификации состоит в том, что новый алгоритм не только группирует объекты по схожести описаний, но и производит отбор полезных признаков для улучшения качества группирования.

Предлагаемый подход

Модифицированное правило обучения Хеббиана можно представить в виде следующего алгоритма:

1) инициализируется нейрон (в начале обучения нейрон зависит только от одного признака, который выбирается случайным образом);
2) для каждого объекта из обучающей выборки выполняются шаги 2.1-2.1;
    2.1) вычисляется активность нейрона (чем меньше отличие между нейроном и входным сигналом по отобранным признакам, тем выше активность нейрона);
    2.2) производится коррекция весового вектора только в том случае, если нейрон активен (нейрон ассоциируется с теми признаками, значения которых существенно отличаются от среднего значения во время активности нейрона).

Нейрон, функционирующий по модифицированному правилу обучения Хеббиана (НМХ), является элементом ассоциативной памяти. Процесс обучения нейрона состоит в образовании ассоциаций между активностью нейрона и признаками (входными сигналами), но ассоциации не должны образовываться в результате случайного

стечения обстоятельств. По этой причине предлагаются два критерия “не случайности” работы нейрона:

- нейрон не должен активизироваться в том случае, если на вход сети подается сигнал, состоящий из независимых случайных величин;
- активность нейрона не должна ассоциироваться с признаком, который является независимой случайной величиной.

В практической реализации понятие “не должен” означает, что такое событие происходит с очень маленькой вероятностью и является практически невозможным. Указанные два критерия были положены в основу модифицированного алгоритма обучения нейрона. *Цель обучения* нейрона - сгруппировать объекты и признаки для достижения максимальной не случайности группирования. *Правило активации* нейрона имеет стандартный вид:

$$Y = F\left(\frac{\sum_{i=1}^{N} w_i \cdot x_i - w_0}{c}\right),$$

где W - весовой вектор нейрона; Y - активность нейрона; w_0 - порог активации нейрона; C - нечеткость порога активации нейрона. $F(x)$ вычисляется по формуле: $F(x) = \frac{1}{1+e^{-x}}$. *Правило обучения* состоит в коррекции весового вектора нейрона w, средней активности нейрона A, порога активации нейрона w_0 и нечеткости порога активации нейрона C. Для обучения нейрона необходимо знать средние значения входных сигналов - вектора P, скорости обучения λ и уровня не случайности Q, который обычно равен: $Q = \Phi^{(-1)}(0.9999) = 4$, где $\Phi(x)$ - интеграл вероятностей для нормального распределения. *Коррекция средней активности* нейрона производится по формуле: $A = A \cdot (1-\lambda) + \lambda \cdot Y$, где Y - активность нейрона; A - средняя активность нейрона.

Коррекция весового вектора нейрона W производится по формуле:

$$w_i = G_i\left(G_i^{(-1)}\left(w_i \cdot (1-\lambda \cdot Y)\right) + \lambda \cdot Y \cdot x_i\right) \quad i = \overline{1,N},$$

где Y - активность нейрона; G вычисляется по формуле:

$$G_i(x) = F\left(\frac{x - p_i}{\sigma_i} - Q\right) - F\left(\frac{p_i - x}{\sigma_i} - Q\right);$$

$$F(x) = \frac{1}{1+e^{-x}}; \quad \sigma_i = \sqrt{\lambda \cdot A \cdot p_i \cdot (1 - p_i)},$$

где w - весовой вектор нейрона; A - средняя активность нейрона.

После коррекции весового вектора нейрона корректируются порог активации нейрона w_0 и нечеткость порога активации нейрона c:

$$c = \sqrt{\sum_{i=1}^{N} w_i^{\,2} \cdot p_i \cdot (1 - p_i)}; \quad w_0 = Min\left(\sum_{i=1}^{N} w_i p_i + Q \cdot c, \sum_{1<i<N}^{w_i>0} w_i\right).$$

Эксперименты и анализ результатов

Приведем описание алгоритма, генерирующего изображения, которые будут использоваться в качестве обучающей выборки в серии машинных экспериментов. Этот алгоритм генерирует изображения, состоящие из вертикальных и горизонтальных линий. Имеется 2N потенциально возможных линий, из которых N вертикальных и N горизонтальных. Каждая вертикальная линия есть прямоугольник размером N точек по высоте и 2 точки по ширине, а горизонтальная - прямоугольник размером 2 точки по высоте и N точек по ширине. Каждая линия состоит из двух рядом стоящих черной и белой полосы. Очевидно, что линии накладываются друг на друга. Каждая точка одновременно принадлежит как к вертикальным линиям, так и к горизонтальным. По этой причине ничего нельзя сказать о классе изображения по одной точке в отдельности. Это обстоятельство приводит к необходимости восприятия каждой точки в контексте с другими точками. Алгоритм состоит из трех шагов:

1) случайным образом определяется класс изображения (оба класса равновероятны);
2) выбираются те линии, которые должны быть нарисованы на изображении, используя граф условных вероятностей, показанный на рис.3;

3) каждая выбранная линия предопределяет подчиненную ей точку с условной вероятностью *0.78*, однако в целом черные и белые точки появляются равновероятно.

Данный алгоритм реализуется с помощью генератора случайных чисел. Черные точки кодируются единицами, а белые - нулями.

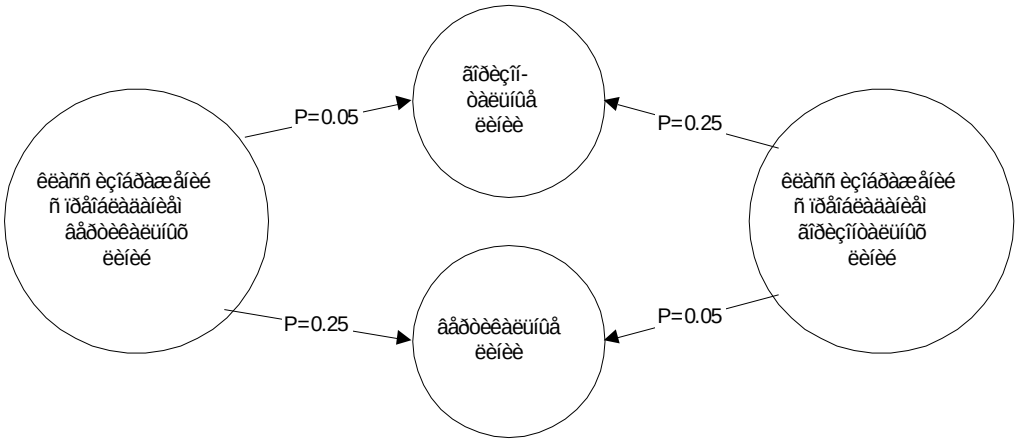

Рис. 3. Схема зависимости между классом изображения и типом линии

Сеть должна была научиться определять класс изображения при условии, что правильный ответ был известен только в 1% случаев. Данное условие обосновано тем, что при принятии решений на практике правильное решение известно значительно реже, чем описание задачи. На рис. 4 показаны типичные изображения из обучающей выборки. Изображения с преобладанием вертикальной штриховки должны быть отнесены к первому классу, а с преобладанием горизонтальной - ко второму.

1. Эксперимент - динамика изменения активности нейрона

В процессе обучения нейронной сети в каждом нейроне формируется реакция нейрона на входные данные. При построении правила обучения нейрона предполагалось, что искусственный нейрон будет настраиваться на одно из устойчивых сочетаний входных сигналов (например, в эксперименте с двумя классами изображений такими устойчивыми сочетаниями были как вертикальные, так и горизонтальные линии, или группы линий). Для проверки описанной гипотезы исследовалось формирование реакции нейрона на предъявление ему определенных линий в процессе обучения сети. Для этой цели был выбран один из нейронов. После каждого шага обучения

вычислялась средняя активность выбранного нейрона при появлении каждой из линий. На рис.6 показано, как нейрон постепенно настраивался на 17-ую линию. При появлении данной линии постепенно возрастала средняя активность нейрона. Можно заметить, что в начале обучения средняя активность нейрона была высока при появлении 17-ой и 38-ой линий. В ходе обучения ситуация менялась. Нейрон переставал активизироваться при появлении 38-ой линии. В результате обучения нейрон стал активизироваться только при появлении 17-ой линии.

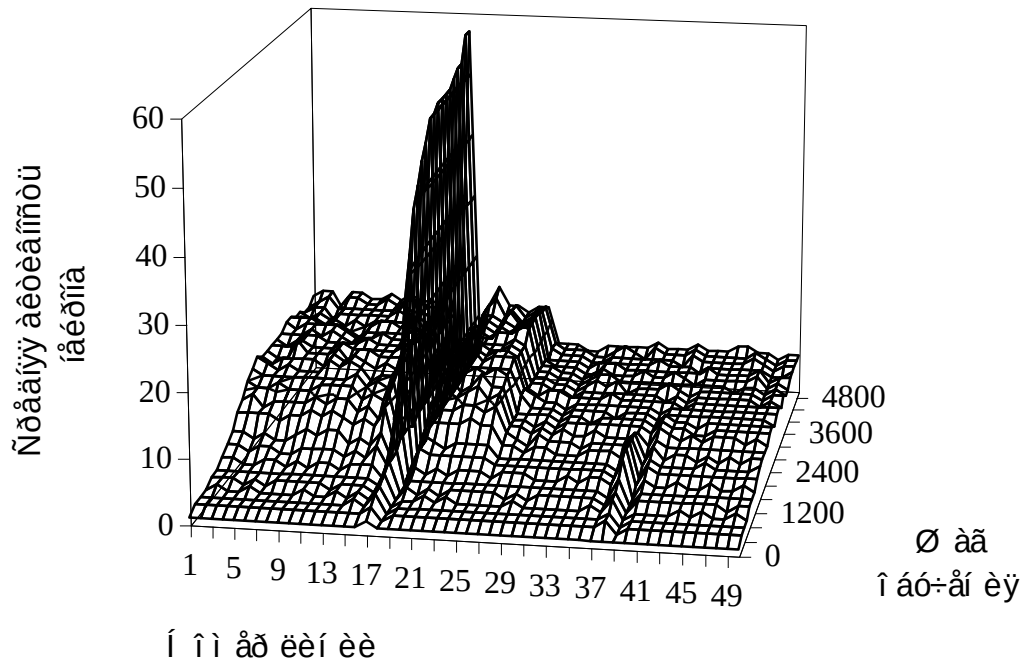

Рис.6. Изменения средней активности нейрона в процессе обучения

2. Эксперимент - кластеризация в нейронах сети Кохонена и в НМХ

Первоначальная постановка задачи была связана со сложностями обучения нейронной сети Кохонена, а именно с тем, что при группировании объектов данных в пространстве большой размерности возникает необходимость группировать объекты в подпространствах данного пространства. Сеть Кохонена не имеет возможности переходить от пространства к подпространствам при группировании объектов, т.к. не

ранжирует признаки по степени их важности. Однако модифицированное правило Хеббиана одинаково хорошо справляется с задачей группирования в обоих случаях. В эксперименте предполагалось проверить эти рассуждения.

Сначала исходное пространство объектов было сформировано таким образом, чтобы объекты группировались во всем пространстве признаков. На вход сети подавались изображения букв “A” и “B”, на которые в последствии накладывался шум. В результате обучения нейронной сети Кохонена сформировались весовые векторы нейронов. Наиболее типичные из них представлены на рис.8.

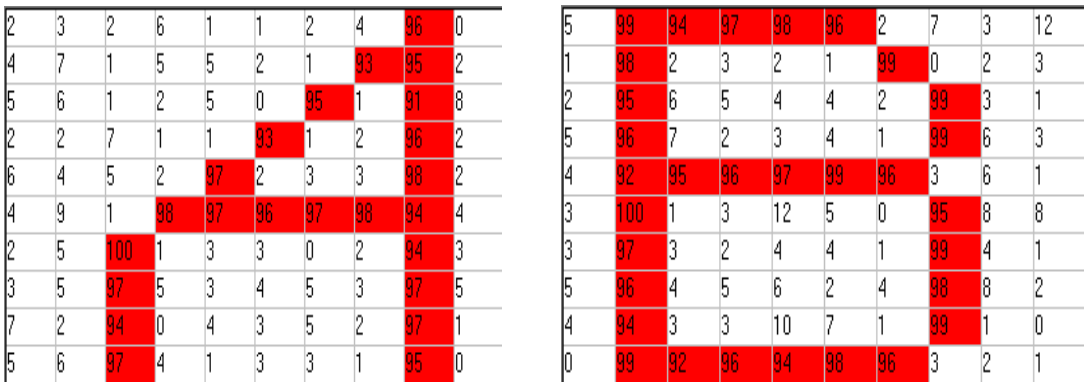

Рис. 8. Весовые векторы некоторых нейронов сети Кохонена. Некоторые нейроны реагируют на появление буквы “A”, а другие на появление буквы “B”

В результате обучения некоторые нейроны стали реагировать на появление буквы “A”, а другие - на появление буквы “B”. Таким образом, карта Кохонена достаточно хорошо справилась с задачей группирования. Сеть, состоящая из НМХ, решала ту же самую задачу. На рис.9 представлены наиболее типичные весовые векторы нейронов.

Рис. 9. Весовые векторы НМХ. Некоторые нейроны реагируют на появление буквы “A”, а другие на появление буквы “B”

Очевидно, что результаты обучения нейронных сетей оказались схожими. Таким образом, подтвердилась гипотеза о том, что сеть, состоящая из НМХ, группирует объекты не хуже сети Кохонена в том случае, когда не нужно ранжировать признаки по степени важности. Во второй части данного эксперимента исходное пространство объектов было сформировано таким образом, чтобы объекты плохо группировались во всем пространстве признаков, и было необходимо ранжировать признаки по степени их важности. Обучающая выборка состояла из изображений двух классов: вертикальных и горизонтальных линий.

В результате обучения нейронной сети Кохонена сформировались весовые векторы нейронов. На рис.11. представлен весовой вектор типичного нейрона. Можно заметить, что в результате обучения нейроны стали реагировать на появление некоторого сочетания вертикальных и горизонтальных линий. По этой причине для нормальной работы карты Кохонена необходимо, чтобы для каждого сочетания вертикальных и горизонтальных линий был зарезервирован отдельный кластер, т.е. искусственный нейрон. Это обстоятельство может создать много проблем при реализации нейронного алгоритма, т.к. необходимое количество кластеров будет экспоненциально расти в зависимости от количества линий (для изображения N на N точек больше, чем 10^{15}). Как правило, при реализации алгоритма количество зарезервированных кластеров сравнительно небольшое, поэтому реакция сети в большинстве случаев будет случайной, т.к. те сочетания линий, на которые настроятся нейроны будут встречаться достаточно редко.

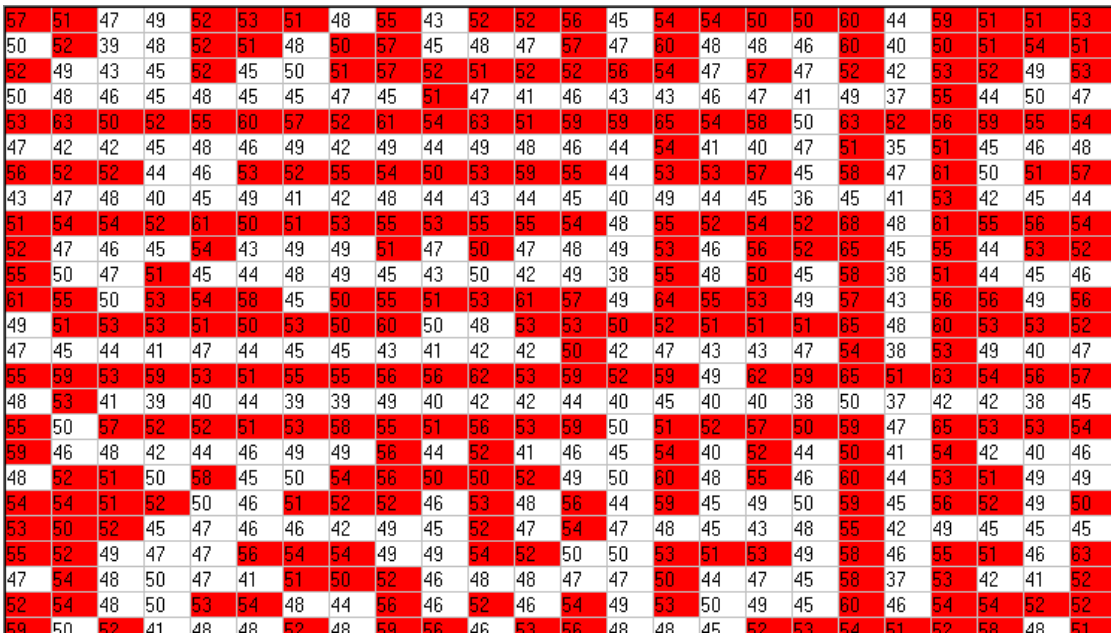

Рис. 11. Весовые векторы некоторых нейронов сети Кохонена после обучения на изображениях, состоящих из вертикальных и горизонтальных линий

Сеть, состоящая из НМХ, решала ту же самую задачу. На рис. 12 представлен весовой вектор типичного нейрона после обучения. Каждый из НМХ настраивается на некоторую линию, поэтому в большинстве случаев реакция такой сети будет неслучайна, т.к. каждая из линий часто встречается на изображениях. Если задача распознавания класса изображения, поставленная перед сетью, будет выражаться через линии или другие устойчивые сочетания точек (например, задача распознавания класса изображений, в которых вертикальных линий больше, чем горизонтальных), то задача линейно разрешится.

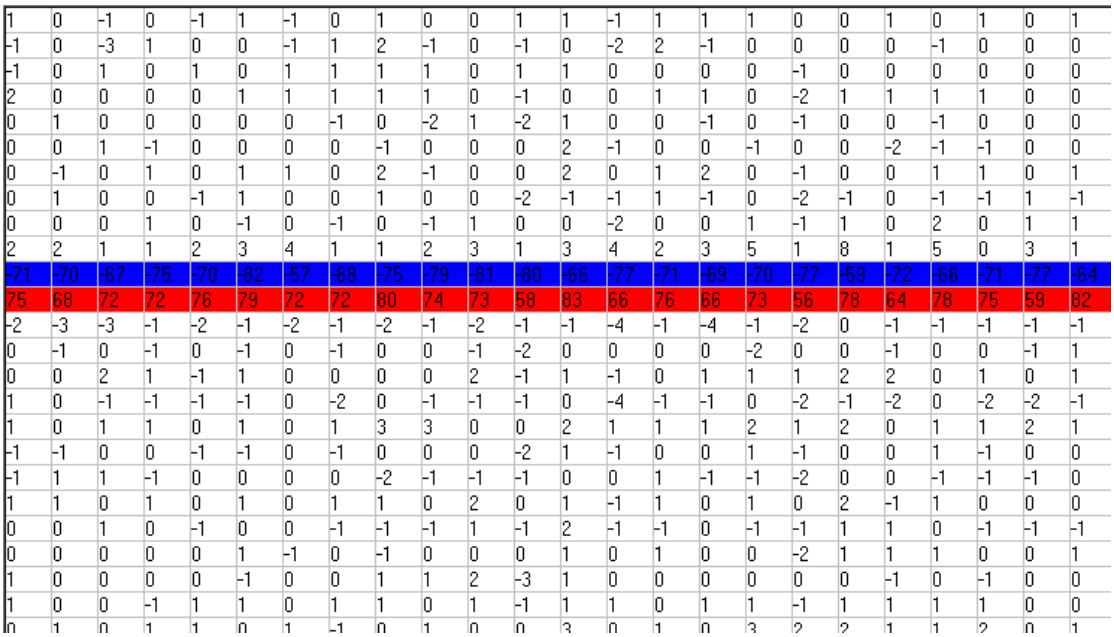

Рис. 12. Весовой вектор НМХ, который настроился на одну из горизонтальных линий

Очевидно, что результаты обучения нейронных сетей обеих архитектур оказались различными. Различие состоит в том, что сеть Кохонена не имеет возможности переходить от пространства к подпространствам при группировании объектов, а НМХ одинаково хорошо справляется с задачей группирования в обоих случаях. Таким образом, подтвердилась рабочая гипотеза.

3. Эксперимент - распознавание класса изображения

Как уже отмечалось, использовались изображения с линиями двух классов:

- класс изображений с преобладанием вертикальных линий;
- класс изображений с преобладанием горизонтальных линий.

Для сравнения точности распознавания класса изображения сети Кохонена с точностью распознавания сети, состоящей из НМХ, проводился машинный эксперимент. На вход нейронных сетей подавались растровые изображения двух классов, размером N на N точек. Обучающая выборка состояла из 5000 изображений, и правильный ответ был известен в 50 случаях. Архитектуры сетей состояли из двух слоев нейронов (см. рис. 14):

- скрытый слой - кластеризующие нейроны;
- внешний слой - один перцептрон.

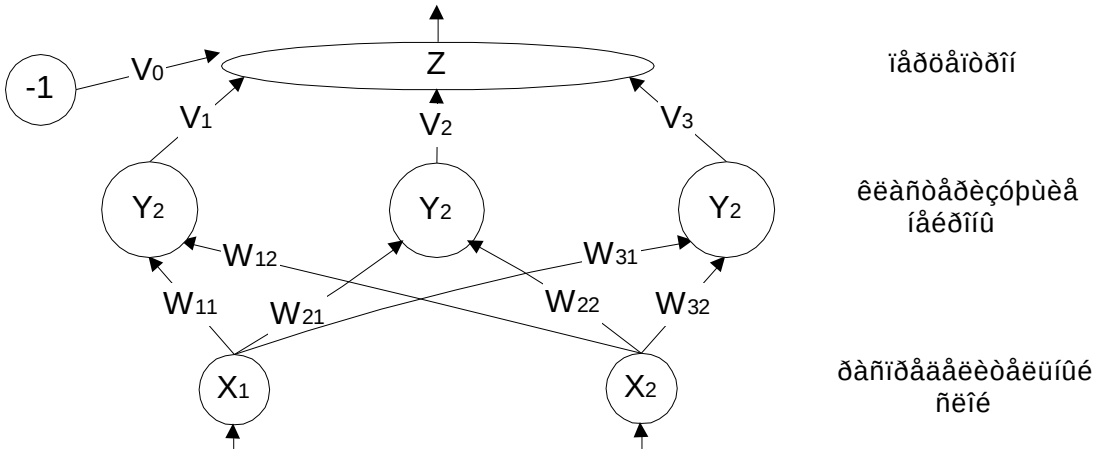

Рис. 14. Архитектура нейронной сети

Сеть, состоящая из НМХ, имела 20 нейронов в скрытом слое. Карта Кохонена была выбрана размером 5 на 5 нейронов. Обучение происходило в два этапа: *кластеризация и обучение с учителем*.

- На первом этапе кластеризации сеть Кохонена и сеть, состоящая из НМХ, обучались на всей выборке из 5000 изображений. Эти алгоритмы имеют возможность обучаться без сигнала от учителя. Обучение персептрона во внешнем слое не происходило.
- На втором этапе фиксировался слой кластеризующих нейронов, обучение нейронов этого слоя не происходило. Информация о классе была известна для 50 изображений. Эта информация использовалась для обучения перцептрона.

После каждого шага обучения нейронных сетей с учителем (полного прохода по всей выборке из 50 изображений) вычислялась ошибка классификации - процент неправильно распознанных изображений. Для этой цели использовалась выборка для тестирования, состоящая из 200 изображений, которые не использовались при обучении. На рис. 15 показано изменение ошибки в процессе обучения с учителем в зависимости от N - количества линий.

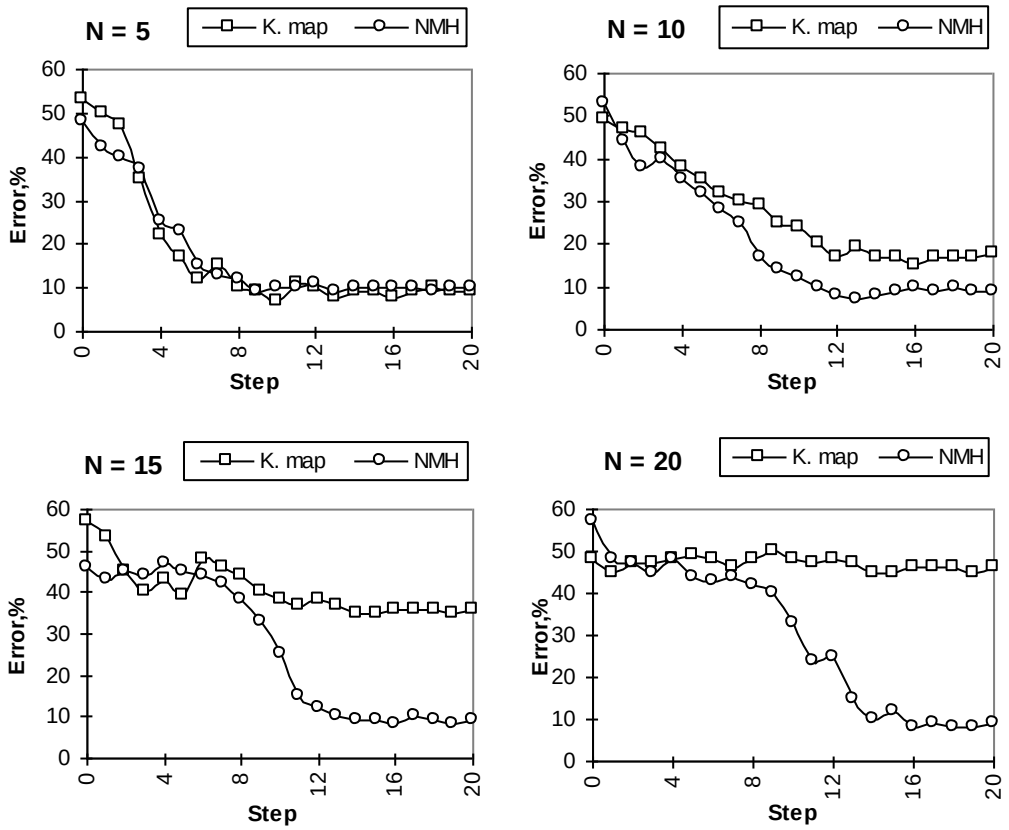

Рис. 15. Изменение ошибки в процессе обучения нейронных сетей с учителем

Нетрудно заметить, что нейронная сеть, построенная на базе НМХ, достигает значительно более высокой точности при распознавании класса изображения при увеличении количества линий.

Выводы

- Выбраны принципы модификации искусственного нейрона и критерий отбора полезных признаков для увеличения качества кластеризации - неслучайные ассоциации.
- На базе искусственного нейрона был получен и реализован алгоритм, ведущий отбор полезных признаков при кластеризации.
- Проведена последовательность машинных экспериментов, для сравнения работоспособности нововведенного алгоритма с сетью Кохонена.
- В экспериментах нейронная сеть, основанная на НМХ, показала значительное превосходство в точности и эффективности работы по сравнению с картой Кохонена.

- В результате данного исследования можно сделать вывод, что отбор полезных признаков во время кластеризации играет важную роль.

Список литературы

1. Дж. Вэн Райзин. (1980). *Классификация и кластер,* Мир, Москва.
2. Хант Э. (1978). *Искусственный интеллект*, Мир, Москва.
3. Eberhart R., Simpson P. & Dobbins R. (1996). *Computational Intelligence PC Tools. AP Professional*, Academic Press Inc.
4. Fausett L. (1994). *Fundamentals of Neural Networks, Architectures, Algorithms and Applications.* Prentice Hall Intern. Inc.
5. Allinson, N. M., Taylor, J. G., Mannion & C. L. T. (1992). Self-Organising Neural Maps and their Applications. In: *Theory and Applications of Neural Networks.* Springer-Verlag, London, p. 101, 101-118.